Capstone Projects

The theme of the capstone projects revolves around ethics, data, and society. As data scientists, ethics are an important part of our everyday lives, yet we rarely discuss the ethical implications of our actions and our inactions. Nearly everything that we do with data, from the collection to the cleaning to the analysis to the use of the information therein obtained, can be used to impact our lives, those around us, and even society in general. Consequently, we must consistently act and perform our roles as data scientists in an ethical manner.

In the capstone project, you will perform research in an area focused on the intersection between data and society. Broad research topic domains are given from which you will identify, in collaboration with the course professors and the project sponsor(s), a specific research problem that you will address for your capstone project. The general themes involved in each of these topics include ethics, data, and society and the interactions between these three. Your research on your specific problem will address all three of these themes. *Specifically, as part of this research, you will perform an ethical analysis of the problem, the solution, the impacts of both or any combination of these and include this analysis as one section in your final research paper*. This ethical analysis will be performed regardless of the problem being solved for the capstone project.

While your research will be performed largely by you and your research group, your will receive feedback on your work over the course the capstone project. This feedback will come from the professors, the project sponsors, and from your peers in the course. In addition, it is strongly encouraged that you seek out one or more capstone advisors to act as guides for your research, sounding boards for your ideas, or other sources of support. Advisors may be professors, industry researchers or other qualified persons. Your advisors should be in addition to any project sponsor(s) that you may be working with.

The documentation of your research and findings will consist of a research paper to be published in the *SMU Data Science Review* journal, a poster presentation of your work to be presented at the conference, and a presentation of your work at the conference. The lasting documentation of your research will be the research paper. The *SMU Data Science Review* is an open access journal published by SMU to act as a searchable and accessible record or your capstone research.

Example Capstone Project Topics

The following list of capstone project topics is an example list from which you may develop a specific target problem for your capstone research. However, this is not an exhaustive list. If you have a problem idea that involves the intersection of ethics, data, and society, then that problem is relevant and a possibility for your capstone research. Please discuss your problem ideas with the course professor before pursuing them too deeply.

Corporate Activities – Corporations perform a wide variety of data acquisition and analysis of their customers. Some of this acquisition and analysis is widely disclosed and intended to provide better services to customers. However, sometimes this activity is not disclosed or is compelled under secrecy rules. What is the impact on society of corporate activities such as scanning emails and monitoring their users both with and without their knowledge?

Unintended Consequences of Location Information – location information is commonly used by advertisers and social media companies to provide contextually aware advertisements and recommendations. However, these recommendations have the potential to violate privacy, for example, by recommending connecting with other people who frequent the same locations. What are the privacy, societal and ethical implications of using location information for advertisement and social media applications?

Ethical Implications of Perfect Recall – Much of our lives is being captured on cameras. Each still image or video shows an incomplete view of a portion of our activities. Other portions of our lives are being captured, however imperfectly, by ourselves and others in social media posts. How do digital memories of imperfect data, both implanted and recorded, impact law and society?

Right to be Forgotten – The European Union allows private citizens to request that certain information about them be removed from Internet search engine results. The effect of this removal is to make it very difficult to find the information about an individual, effectively erasing a portion of an individual's history. Public individuals do not have this protection; however, any information "forgotten" while an individual is a private citizen is not automatically "remembered" when they become public individuals. What does the right to be forgotten do to true memory, the collective memory of society and to history?

Privacy Issues in a Cashless Society – A cashless society relies solely upon digital transactions for the purchase of goods and the exchange of wealth. All of these financial transactions, in accordance with standard financial practices and auditing laws and rules, are traceable. When all financial transactions are digital, does the resulting cashless society reduce our privacy?

The Return of HAL – We are currently witnessing the rise of intelligent machines through the use of artificial intelligence, machine learning and advanced analytics techniques. What are the ethical and societal implications of human intelligence in machines?

Stingray and Big Brother – The Stingray cell tower has been widely used by law enforcement to monitor the cell phone activities of large numbers of citizens in the pursuit of criminals utilizing their cell phones for illegal activities. What are the consequences, actual, likely and/or possible, on society and the technologies that we use due to this widespread indiscriminate surveillance?

Wheat from the Chaff – Anonymized data sets are commonly made available for research purposes, particularly for drug trials and other medical and financial research purposes.

However, recent research has shown that when combined with other public data sets, individuals in the anonymized data sets can often be uniquely identified. One approach to counter this type of analysis is to modify the anonymized data. How much can the anonymized data be modified in order to mitigate the identification of individuals in the data set while allowing for the same analysis results of the data to be achieved?

The Rights of People to Their Data – Every person generates huge amounts of data through their everyday lives and activities. All digital activities are monitored and recorded to greater and lesser degrees, leaving at least bread crumbs to indicate the passage of a particular person through the ether. Even travel through the physical world leaves electronic bread crumbs through the numerous cameras and electronic devices that travel with and around us. Most of these digital bread crumbs (some of which are large loaves of bread) collected about an individual are not under that individual's control, are not collected by that individual, and are not accessible by that individual. Given the ever increasing digital life led by everyone, what are an individual's rights to the data collected about him/her? What laws are in place to protect an individual? What laws protect the collector of the data? Should these laws by changed?

How much Privacy do we have Today? – Much of our lives are available online for anyone that cares to look at us. But, how much can someone find out about an individual just from publicly available information? In this project, select a public or semi-public individual, such as a politician, a judge, or a celebrity, and develop a detailed record of their life.

Air Pollution and Increased Death Rates – A number of studies have found a positive correlation between high air pollution and increased mortality. Using recent data, evaluate the impact of air pollution on the daily health (and death) of a major city.

MSDS Capstone, Summer 2017: BiliCam: Regression of Image Features with Neural Networks

Principal Contact: Eric C. Larson, Computer Science and Engineering

Description

This project is one part of a larger, ongoing project that seeks to estimate the level of BiliRubin in the blood of a newborn, typically called jaundice. The jaundice level is predicted from a several pictures of the new the newborn taken from the camera of a smartphone (in our testing, we have used iPhone 5S hardware). While the bulk of the feature processing and machine learning regression has already been developed, we would like to investigate nonlinear methods and more advanced regressions, potentially using deep neural networks. The availability of the data will mostly be in a format that is already processed (i.e., features from the images have already been extracted and the image data in raw form may not be available). The data consists of processed features from multiple images, collected under different distances and lighting conditions. Image data exists for approximately 500 newborns with ground truth blood draw labeling.

Personnel

An ideal team of students might consist of: familiarity working with image data, traditional machine learning methods, and (possibly) neural networks.

Deliverables

I will work with the team to develop a set of concrete deliverables, but ideally students would answers the following during the capstone:

- What is the best cross validated linear model performance in regards to predicting bilirubin from image feature data?
- What feature subsets have statistically non-different performance from the best linear model?
- What neural network architectures can capture bilirubin and how do these compare to traditional linear models?
- Given raw image data, can these results be improved using convolutional neural networks (depends on availability of data and ability of students to become certified in human subjects research)?

MSDS Capstone, Summer/Fall 2017: Bipolar Awareness Prediction via Analysis of Eye-based Landmarks

Principal Contact: Eric C. Larson, Computer Science and Engineering

Description

This project is proposed by the students below and I was asked to help advise the students during their investigation. The problem can be scoped as follows: Manic-depressive disorder is a complicated illness and manifests in different ways for different individuals. One method of detecting manic episodes comes from anecdotal evidence that changes in the eyes reveal when an individual is experiencing mania. In this scenario, pictures could be used to analyze if someone is experiencing a manic episode. These episodes are difficult for someone experiencing them to be aware of—thus an automatic identification system would help increase awareness of mania in the individuals and in loved ones/care givers. To validate this concept, a data collection protocol along with introductory analysis is proposed. This project is exploratory research and will require considerable effort for all students involved.

Personnel

Jessica Wheeler, Sharon Teo, Jean Jecha, and Manjula Kottegoda. Assistance from Julie Fast, a long time researcher in bipolar disorder, will also be contributed. This assistance is given in-kind (without monetary reimbursement).

Deliverables

I will work with the team to develop a set of concrete deliverables, but ideally students would answers the following during the project:

- How can data be collected and managed with ground truth labels for degree go mania being experienced? Does ecological momentary assessment help to capture a reliable ground truth?
- Receive human subjects approval to carry out data collection.
- What methods of analysis of the eyes are critical for capturing manic episodes? Color? Pupil dilation? Gaze patterns? Facial expression?
- Can convolutional networks centered on the eyes of individuals be used to capture manic episodes? How much data is required to make the system reliable?
- Are any features of the eyes generalizable across participants? Or does the prediction process require a calibration protocol?

Capstone, Summer/Fall 2017: Models and Data Structure Comparison

Principle Contact: Bivin Sadler

Description

A common question that keeps popping up in my head is, "When does a Random Forrest perform better than other, more traditional, competing methods?" Specifically, it would be interesting to investigate when random forest classification performs better than logistic regression? In order to research this question, one could simulate data of various complexity (number of features) and various correlation structures. It stands to reason that if you simulate data from a linear model with known parameter estimates and independent features, that logistic regression would work best since it is actually estimating the true structure the data was pulled from. A) Is this true? I often assume that something will be true and/or easy just to find something amazing after a little research. B) What if we add a non-independent correlation structure to the linear model? C) What if generate the data from a a non-linear model … maybe research a known quantity that has a non-linear (and non curvi-linear) relationship with the logit and compare and contrast logistic regression and a random forest model in classifying these data? It would be nice to be able to give practitioners and clients from different fields some guidance as to which method they should use if they know something about the structure of their data.

Note: This project could be spun off in quite a few different directions. Another group could look at differing sample sizes, while other groups could simply pick off a different correlation structure or model complexity that that is commonly found in a different field. In addition, under the same umbrella but in a slightly different direction, a group cold look at continuous responses and contrast linear models (MLR) with random forest regression. Of course, in the end, there are many more models to compare than simply these 4 … groups could compare and contrast nearly any modeling/predictive method with respect to the above framework.

Personnel

An ideal team of students will have familiarity with and interest in classification models including regression models and random forest models.

Deliverables

We will work with the team to develop a set of concrete deliverables to answer the questions above.

MSDS Capstone, Summer/Fall 2017: Cognitive virtual admissions counselor

Principal Contact: Raghuram Srinivas, Computer Science and Engineering

Description

Every day the admissions counsellor get hundreds of calls from prospective students about your program , the MSDS program at SMU. Most prospective students' questions center around the course outline, program costs, the experience of the mode of delivery and the like. These questions and more ,can be addressed a very human like 24/7/365 online student advisory service

This project involves training a Watson/Alexa like system on natural language in the admission advisory domain. The system will be trained to understand a prospects questions and be able to respond based on its training. It will also be able to adapt to newer behavior as it receives feedback from its interactions so it adjusts for the next time the question is asked

Personnel

An ideal team of students comprises of familiarity to work with Natural language system such as IBM Watson or Amazon Lex services. Proficient in programming / scripting languages such as python or nodejs to develop integration services. Familiarity with traditional machine learning skills to detect need for retraining and adaptive learning

Deliverables

We will work with the team and MSDS admissions counsellors to develop the scope and concrete deliverables.

MSDS Capstone, Fall 2017: OpenCycle: forecasting ovulation for family planning through

Principal Contact: Eric C. Larson, Computer Science and Engineering

Description

Predicting fertility has a number of uses in family planning including avoiding pregnancy and assisting couples in becoming pregnant. We intend to create a data-driven family planning technique that relies on simple, at-home measurements collected via a mobile phone. Conventional fertility assessment relies on a rule known as "three-over-six": when a woman's minimum basal body temperature (BBT) over the past three days has exceeded the maximum for the six days before, we can be reasonably sure that ovulation occurred three days ago. While informative, this technique can only inform couples that ovulation has already occurred. A more useful algorithm would inform couples in real time about fertility and forecast the most fertile times.

We have acquired a large dataset of recorded menstrual cycle data with which we can inform the design and evaluate various forecasting algorithms in their ability to accurately predict ovulation. This gives more time to families who seek to have children - and limits risk for those who do not. We will explore time series models, deep learning, and various cross-validation techniques to ensure statistical validity of our results. As the data are already collected, we will not need to involve human subjects.

Personnel

An ideal team of students might consist of: familiarity working with mixed time series data and categorical data, traditional machine learning methods.

Deliverables

I will work with the team to develop a set of concrete deliverables, but ideally students would have answers to the following:

- How accurately can we forecast an individual's day of ovulation. What are our false positive and negative rates?
- How many days' data do we need to predict ovulation?
- Can we predict that ovulation has already occurred?
- What are the relative importances of an individual's personal history compared to sub-group's data in prediction?
- How much personal history do we need to forecast one person's ovulation?
- How do we handle anomalies? That is, people whose cycle lengths are abnormal? Do we remove their data from training? Do we attempt to predict their ovulation date?
- Which personal demographic details are beneficial to predicting ovulation? E.g. age, race.
- How sensitive is the model to missing BBT measurements? That is, can the algorithm recover when one or two BBT measurements are missing?

MSDS Capstone, Summer/Fall 2017: Smart Infrastructure: Detecting and Mapping Dallas Infrastructure

Principal Contact: Eric C. Larson and Raghuram Srinivas, Computer Science and Engineering

Description

The city of Dallas is a vibrant and a major economic center of the North Central Texas region. Like every growing city, Dallas faces challenges such as severe weather events and disease outbreaks that have differing effects on different areas of the city, many times related to economic disparity. In this context there exist several research questions:

- What are the spatial relationships among indicators of social, ecological, and engineered vulnerability in urban settings?
- Are existing and proposed growth initiative projects sited in location that are most beneficial to reducing vulnerability?

The city of Dallas is partnering with several research and educational institutions to address these questions, starting with the assessments of the adequacy of the infrastructure such as quality of sidewalks, waterlines, sewer lines, streetlight, transportation , cell phone and internet coverage. The initial assessments will investigate the effect of infrastructure on social vulnerabilities of residents (especially in terms of health and safety).

In this context we propose 2 projects :

Project 1 (starting in Summer): Side Walk Detection:

There is a need to assess and grade the quality sidewalks in the city of Dallas. This project aims to develop systemic methods to detect the presence of sidewalks via image data, then map the data to a given geolocation. Finally, we wish to grade the quality of the sidewalk automatically.

Project 2 (starting in Fall): Neighborhood Safety Vulnerability indexing

This project aims to study and detect relationships between the city's crime data and property valuation.
https://www.dallasopendata.com/Public-Safety/Police-Incidents/tbnj-w5hb
https://www.zillow.com/howto/api/HomeValuationAPIOverview.htm

Personnel

An ideal team of students might consist of: familiarity working with image data, traditional machine learning methods, and neural networks. Familiarity with GIS databases is also recommended (or desire to learn quickly).

Deliverables

We will work with the team to develop a set of concrete deliverables based upon the research questions above.